Project1: Tweet Analysis

Social media has gained immense popularity with marketing teams, and Twitter is an effective tool for a company to get people excited about its products. Twitter makes it easy to engage users and communicate directly with them, and in turn, users can provide word-of-mouth marketing for companies by discussing the products. Given limited resources, and knowing we may not be able to talk to everyone we want to target directly, marketing departments can be more efficient by being selective about whom we reach out to.

In this project, we'll learn how we can use Apache Flume, Apache HDFS, Apache Oozie, and Apache Hive to design an end-to-end data pipeline that will enable us to analyze Twitter data. The Twitter Streaming API outputs tweets in a JSON format which can be arbitrarily complex. Following tasks need to be accomplished in order:

MileStone1: Analyze Twitter data to understand the contents and structure

MileStone2: Define flume configuration to gather streaming twitter data into hdfs. Verify the correctness.

MileStone3: Create an external HIVE partitioned table that points to the ingested twitter data. Partition key must be (date and hour)

MileStone4: Create an oozie job that adds a partition containing last hour's worth of data into Hive and instruct the workflow to occur every hour.

MileStone5: Develop HIVE queries to answer the following questions:

a) Which Twitter users get the most retweets i.e., who is influential within our industry?

(The mechanics of Twitter may help you to understand the above question. A user – let's call him Joe – follows a set of people, and has a set of followers. When Joe sends an update out, that update is seen by all of his followers. Joe can also retweet other users' updates. A retweet is a repost of an update, much like you might forward an email. If Joe sees a tweet from Sue, and retweets it, all of Joe's followers see Sue's tweet, even if they don't follow Sue. Through retweets, messages can get passed much further than just the followers of the person who sent the original tweet. Knowing that, we can try to engage users whose updates

tend to generate lots of retweets. Since Twitter tracks retweet counts for all tweets, we can find the users we're looking for by analyzing Twitter data.)

b) Which time zones are the most active per day?

c) Which were the most common hashtags?

d) Find the tweet counts per user?

MileStone6: Create a temporary HIVE table that contains the result of last query and export the contents of that table into corresponding relational table in MySQL. (You have to create corresponding table in MySQL before export)